GenEst Statistical Models

10 October 2018

By Daniel Dalthorp, Lisa Madsen, Manuela Huso, Robert Wolpert, Paul Rabie, Jared Studyvin, Juniper Simonis, and Jeffrey Mintz

Contents

Abbreviations

A	$X \times nsim$ array of simulated arrival intervals
A_j	event that the carcass arrived in interval j (between times t_{j-1} and t_j)
g, $\hat{g}$	carcass detection probability, estimated carcass detection probability for carcasses arriving in searched areas
CDF	cumulative distribution function
CP	carcass persistence
dwp_u	density-weighted proportion or the fraction of carcasses falling at search unit u that fall within the searched area
f	sampling fraction or the fraction of carcasses at a site that fall at the units selected for search
E[X]	expectation of the random variable, X
k	fractional change in searcher efficiency with each successive search after carcass arrival
I_j	$t_j - t_{j-1} =$ length of search interval j
M, $\widehat{M}$	mortality, estimated mortality
$nsearch$	number of carcass search days during the monitored period
$nsim$	number of simulation replicates for accounting for variance
O_i	event that the carcass is observed in search i (at time t_i)
p	searcher efficiency for carcasses in the first search after carcass arrival
$S(t)$	carcass persistence distribution = probability carcass persists ≥ t days after carcass arrival
SE	searcher efficiency
t_j	time (in days) of the jth search from the start of monitoring
X	number of carcasses observed
IQR	inter-quartile range (25th through 75th percentiles)

GenEst Statistical Models

1 Introduction

GenEst is a suite of statistical models and software tools for generalized mortality estimation. It was specifically designed for estimating the number of bird and bat fatalities at solar and wind power facilities, but both the software (Dalthorp et al. 2018) and the underlying statistical models are general enough to be useful in a variety of situations to estimate the size of open populations when detection probabilities and search coverages are less than 1. In this report, we outline the statistical models and data structures underlying the estimator. The models are numerous, complex, and intricately interwoven. Discussion begins with broad, high-level overviews of the general models. The lower-level technical details are then gradually added in. Broader and less technical discussions on the general context and applications of the models and the use of the software can be found in the software user guide (Simonis et al. 2018), vignettes bundled with the software, and the help files within the software itself.

At its core, GenEst is an elaboration of a binomial probability model $X \sim \text{binomial}(M, g)$, where X is the observed number of carcasses and g is the detection probability. If g is known, then $\widehat{M} = X/g$ is an unbiased estimator for M and the sampling variance of X is the only source of uncertainty about M. In a slightly more complicated scenario in which total mortality is split into two groups, A and B, then $\widehat{M} = \widehat{M}_A + \widehat{M}_B = \frac{X_A}{g_A} + \frac{X_B}{g_B}$ is unbiased for M. GenEst makes extensive use of this simple idea of splitting the carcass observation data into distinct subgroups, estimating mortality in each subgroup, and combining subgroups into larger groups by summing. A number of technical difficulties must be overcome to make this simple idea work as a complete estimator that produces accurate confidence intervals, including accounting for the dependence of g on the time of carcass arrival, estimating g as a function of covariates, characterizing the uncertainty in $\hat{g}$, accounting for correlation structure of the $\hat{g}$'s among various subgroups, accounting for uncertainty in estimating M given X and g, and accounting for unsearched area. GenEst provides solutions for each of these difficulties. The technical details are lengthy but are explained in full in sections 2-8.

2 Splitting Mortality Estimates by Carcass and Recombining into Subgroups

For each carcass, $x = 1, \ldots, X$, that is discovered in carcass surveys, its contribution to the estimated total mortality is the reciprocal of its inclusion probability, $1/g_x$, but with two types of uncertainty that we account for via parametric bootstrapping (more specifically, simulating from the asymptotic distributions of the maximum likelihood estimators of parameters) with $nsim$ repetitions: (1) uncertainty associated with estimating g_x (sections 7-8), and (2) the uncertainty associated with estimating $M_x|g_x$ (section 5). The result is stored in an $X \times nsim$ matrix of carcass contributions to the estimated total mortality

$$\widehat{\mathbf{M}} = \begin{bmatrix} \widehat{m}_{1,1} & \widehat{m}_{1,2} & \cdots & \widehat{m}_{1,nsim} \\ \widehat{m}_{2,1} & \widehat{m}_{2,1} & \cdots & \widehat{m}_{2,nsim} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{m}_{X,1} & \widehat{m}_{X,2} & \cdots & \widehat{m}_{X,nsim} \end{bmatrix}.$$

Each column of $\widehat{\mathbf{M}}$ represents a single realization of the simulated contributions of all the discovered carcasses to the estimated total mortality. Uncertainty is captured in the distinctions among columns.

Mortality estimation by carcass group is accomplished by subsetting the rows and then taking column sums. For example, total mortality (no subsetting) is estimated by taking column sums over all the carcasses to create a vector of estimates, $\widehat{M}_{total} = (\sum_x \widehat{m}_{x,1}, \sum_x \widehat{m}_{x,2}, \ldots, \sum_x \widehat{m}_{x,nsim})$, from a parametric bootstrap from the sampling distribution of $\widehat{M}$. The $\widehat{\mathbf{M}}$ matrix can be subsetted to estimate mortality among desired subgroups in a similar way. For example, to estimate mortality for species B, calculate $\widehat{M}_{\mathrm{B}} = (\sum_{x \in \mathrm{B}} \widehat{m}_{x,1}, \sum_{x \in \mathrm{B}} \widehat{m}_{x,2}, \ldots, \sum_{x \in \mathrm{B}} \widehat{m}_{x,nsim})$, where $x \in \mathrm{B}$ indicates that carcass x belongs to species B. GenEst extracts sample statistics (like median and confidence intervals) from these empirical $\widehat{M}$ vectors to summarize mortality estimates for subgroups or "splits" as defined by the user.

3 Temporal Splits

GenEst can also split mortality estimates by user-specified time intervals or by temporal covariates like season. To split the $\widehat{\mathbf{M}}$ matrix by carcass arrival times, GenEst relies on an $X \times nsim$ matrix of simulated arrival intervals, **A**, to construct an $ntimes \times nsim$ array of mortality estimates by time interval, $\widehat{M}_T$, as:

$$\widehat{\mathbf{M}}_{\boldsymbol{T}} = \begin{pmatrix} \sum_{x \in I_1} \widehat{m}_{x,1} & \cdots & \sum_{x \in I_1} \widehat{m}_{x,nsim} \\ \vdots & \ddots & \vdots \\ \sum_{x \in I_{ntimes}} \widehat{m}_{x,1} & \cdots & \sum_{x \in I_{ntimes}} \widehat{m}_{x,nsim} \end{pmatrix},$$

where $x \in I_j$ indicates that the simulated arrival time of carcass x is within the specified interval, j. If the simulated arrival interval of a carcass x_i intersects two or more of the splits intervals (for example, I_j and I_{j-1}), the carcass's contribution to estimated mortality, $\widehat{m}$, is allocated proportionally to the intersecting splits intervals.

To understand the structure of the **A** matrix, first note that detection probability for a carcass depends in part on its arrival time. For example, if monitoring runs from April 1 through October 31, carcasses arriving near the end of October may be available for discovery for only one or two searches, while carcasses arriving in April may be available for many searches. Also, search conditions may vary with season. However, although carcass discovery times are recorded, arrival times are not known. Discovered carcasses may have arrived sometime after the previous search or in any search interval prior to that. Even though arrival times cannot be known with certainty, GenEst analyzes the search data to derive probability distributions of arrival times for each carcass and, from these, creates an $X \times nsim$

matrix of simulated arrival intervals, $\mathbf{A}$. For example, if $X = 3$ and the carcasses were discovered on searches $i = 1, 3,$ and 12 of the monitoring period and $nsim = 10$. The arrival matrix, $\mathbf{A}$, might look like the following:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 3 & 2 & 3 & 1 & 3 & 3 & 3 & 2 & 1 \\ 12 & 11 & 12 & 10 & 9 & 12 & 12 & 12 & 12 & 11 \end{bmatrix}$$

The simulated arrival intervals for the first carcass are identically 1 because the carcass was discovered on the first seach (at t_1) and is assumed to have arrived in the first search interval, $(t_0, t_1]$. In practice, the assumption that all carcasses discovered in carcass surveys arrived after t_0 is often enforced by disregarding carcasses found in a careful clean-out search at t_0. The second carcass was discovered in search 3. According to the arrival probabilities, the carcass was more likely to have arrived in interval 3 than any other interval, but there is a chance it arrived as early as the first interval, as reflected in the row of simulated arrivals. The third carcass was found on the 12th search and, in theory, could have arrived in any interval prior to its discovery. However, it is highly unlikely that it arrived more than a few intervals prior to the 12th because, if it had, chances are that it would have been removed by scavengers or previously found by searchers.

4 Estimation of Arrival Probabilities

Suppose searches are conducted at times, $t_0, t_1, \ldots, t_{nsearch}$. For each carcass discovered in carcass surveys $i = 1, \ldots nsearch$, define O_i as the event that the carcass was observed during search i (at time t_i) and A_j as the event that the carcass arrived in interval $j = \left(t_{j-1}, t_j\right], j \leq i$. Then, the probability that the carcass arrived in interval j can be estimated as:

$$\Pr(A_j | O_i) = \frac{\Pr(O_i |\, A_j)\Pr(A_j)}{\sum_{j \leq i} \Pr(O_i |\, A_j)\Pr(A_j)}$$

In theory, $\Pr\left(A_j|O_i\right)$ is positive for $j = i, i-1, \ldots, 1$, but, in practice, $\Pr\left(A_j|O_i\right)$ decreases rapidly to 0 with j, so in most cases only the first few terms need to be calculated. The default for the GenEst software is to calculate arrival probabilities for up to 8 intervals prior to discovery, but the R command-line user may override the default in functions `estM` and `estg`. GenEst makes a further assumption that the arrival rate is constant in the search intervals preceding carcass arrival, so

$$\Pr\left(A_j|O_i\right) = \frac{\Pr(O_i|\,A_j)/I_j}{\sum_{j\le i}\Pr(O_i|\,A_j)\Pr\left(A_j\right)/I_j}$$

where $I_j = t_j - t_{j-1} =$ length of interval j.

The estimator is robust to variation in arrival rates among different search intervals provided that there is not an abrupt change in arrival rate from one interval to the next. However, even in these unusual conditions, the potential for bias in estimating total mortality would be small because the same pattern would need to recur with a large fraction of the carcasses. The potential for bias would be further attenuated by high searcher efficiency or low persistence times.

Calculation of $\Pr\left(A_j|O_i\right)$ is based on calculation of $\Pr(O_i|\,A_j)$:

$$\Pr\left(O_i|A_j\right) = \begin{cases} 0, & i < j \\ p\int_{t_{j-1}}^{t_j}\dfrac{S(t_i - t)}{I_j}dt, & i = j \\ \left[\prod_{s=0}^{i-j-1}(1 - pk^s)\right]pk^{i-j}\int_{t_{j-1}}^{t_j}\dfrac{S(t_i - t)}{I_j}dt, & i > j \end{cases}$$

where $S(t)$ is the probability that a carcass persists t days after arrival (section 8), p and k are searcher efficiency parameters (section 7), s (the index on the product) runs from the search interval immediately before carcass discovery to the first search interval of the monitored period. A carcass may arrive in one season and be discovered in a later season after search conditions have changed. To account for this possibility, in the calculation of a carcass's $\Pr\left(O_i|\,A_j\right)$ for a given j, the carcass is assigned the search characteristics appropriate to the assumed arrival interval, j. For example, if a carcass is discovered on the first search in the fall, and search conditions (that is, searcher efficiency and carcass persistence parameters) change from summer to fall, then $\Pr(O_i|\ A_i)$ is calculated with the fall SE and CP parameters and $\Pr(O_i|\ A_{i-1})$ is calculated with the summer parameters.

Arrival-interval probabilities, $\Pr\left(A_j|\ O_i\right)$, are calculated for each carcass and search interval, and an integer-valued $X \times nsim$ matrix of simulated arrival intervals, **A**, is constructed. Each column represents the arrival interval of each carcass in a simulated set of carcass arrivals. In other words, each column represents simulated arrival intervals of the set of observed carcasses over the course of one field season or simulated "year". As mentioned previously, carcass detection probabilities depend on arrival times, which are given in the arrival matrix **A**. For each simulated year (column in **A**), the detection probability for each carcass is estimated as

$$\hat{g} = \Pr\left(O|A_j\right) = \sum\nolimits_i \Pr\left(O_i|\ A_j\right)$$

to construct an $X \times nsim$ matrix, $\hat{\mathbf{g}}$, of carcass detection probabilities that are applicable to the carcass at the time of carcass arrival. For example, $\hat{\mathbf{g}}_{x,j}$ is a simulated inclusion probability of carcass x, assuming it arrived in interval $\mathbf{A}_{x,j}$.

5 Uncertainty in Estimating $\boldsymbol{M}|(\boldsymbol{X},\boldsymbol{g})$

The uncertainty in estimating g_x for carcass x is captured in $\hat{\mathbf{g}}_{x,\cdot} = (\hat{g}_{x,1}, \dots, \hat{g}_{x,nsim})$, but there remains a great deal of uncertainty about the true mortality, M, even if g is known, due to the sampling variation in X or the uncertainty in estimating M. GenEst accounts for that uncertainty in a novel way that lends itself well to the splits framework. The process is to estimate the sampling variation in X, rescale the estimated variation to $\widehat{M}$, and adjust for bias (Madsen et al. 2018).

We can assume $X \sim \text{binomial}(M, g)$, but, unfortunately, we don't know M. We can, however, estimate M by X/g and (almost) define a new random variable $\tilde{X} \sim \text{binomial}(X/g, g)$ to account for the variation in $X|(M,g)$. The uncertainty in estimating M would then be accounted for in $\widehat{M} = \tilde{X}/g$. However, X/g need not be an integer and cannot serve as the index for a binomial random variable. Instead, we define $\tilde{X} \sim \mathrm{cbinom}(X/g, g)$, where cbinom is a continuous generalization of the standard binomial distribution (Ilienko 2013), which is implemented in the R package `cbinom` (Dalthorp 2018). The continuous binomial spreads the binomial distribution's probability mass on each integer k to the interval $[k, k+1)$, so the distribution is a smoothed version of the binomial but shifted slightly to the right. The mean of the continuous binomial distribution with parameters m and g is $\int_0^{m+1}[1 - F(x; m, g)]dx$, where F is the CDF of the continuous binomial:

$$F(x; m, g) = \begin{cases} 0, & x \le 0 \\ \dfrac{B(x, m+1-x, g)}{B(x, m+1-x, 0)}, & 0 < x \le m+1 \\ 1, & x > m+1 \end{cases}$$

and

$$B(x, y, g) = \int_g^1 t^{x-1}(1-t)^{y-1} dt$$

is the incomplete beta function. The expectation of a random variable $\tilde{X} \sim \mathrm{cbinom}(m, g)$ is

$$\mathrm{E}[\tilde{X}] = \int_0^{m+1} [1 - F(x; m, g)]dx$$

which is approximately $mg + 0.5$ and can be calculated numerically. Since the expected value of a binomial random variable $X \sim \mathrm{binomial}(m, g)$ is mg, $\tilde{X}$ exceeds X by about 0.5 on average. Thus, $\tilde{X}/g$ would be biased for estimating M, so we subtract off the bias before dividing by g:

$$\widehat{M} = \frac{\left[\tilde{X} - \left(\mathrm{E}\left([\tilde{X}]\right) - x\right)\right]}{g}.$$

Because this estimator is unbiased, mortality estimates can be summed. For example, if $\widehat{M}_A$ is the estimated mortality in area A and $\widehat{M}_B$ is the estimated mortality in area B, then $\widehat{M}_{total} = \widehat{M}_A + \widehat{M}_B$ is an unbiased estimator for the two areas combined. There is nothing special about "area" here; A and B could represent times, species, carcass sizes, search teams, turbine types, or other variables. GenEst takes this idea to the limit and builds the $X \times nsim$ matrix, $\widehat{\mathbf{M}}$, in which each carcass represents its own "area". Users then define how they wish to split the overall mortality into summary groups. In the simplest case, total mortality is estimated as the sum of each carcass's contribution to the total, that is, the column sums of the $\widehat{\mathbf{M}}$ matrix.

6 Accounting for Unsearched Area

Unless the user is interested in estimating mortality strictly in the searched area, $\hat{\mathbf{g}}$ values must be adjusted to account for unsearched area. In practice, some carcasses are likely to fall outside the searched area in a given unit (e.g., in an unsearched part of a search ring at a solar power tower facility or beyond the search radius at a wind turbine). In addition, there may be units at a site that are not searched at all. For each search unit, u, the expected fraction of carcasses that are killed at the unit but fall outside the searched area is the *density-weighted proportion* or dwp_u (Huso and Dalthorp, 2014). The expected fraction of carcasses arriving at the units searched at a site is the *sampling fraction* or *f*. For example, if units at a wind facility are the individual turbines, then f would be the fraction of turbines surveyed. Thus, the contributions of the respective carcasses to the estimate of total mortality are then summarized in a $X \times nsim$ array as $\frac{1}{\mathbf{dwp} \cdot \hat{\mathbf{g}} \cdot f}$, where $\mathbf{dwp} = \mathrm{diag}(dwp)$ is the diagonal matrix of dwp values associated with the carcassess. For example, if carcass 1 was found at unit 17 and was a medium-sized bird, then $\mathbf{dwp}_{1,1}$ would be the dwp for medium-sized birds at unit 17. $\widehat{\mathbf{M}}$ would then be calculated based on this adjusted $\hat{\mathbf{g}}_* = \mathbf{dwp} \cdot \hat{\mathbf{g}} \cdot f$.

7 Searcher Efficiency

Searcher efficiency and carcass persistence (section 8) are estimated from field trials where a known number of carcasses are place and their fate (found or scavenged) is recorded. Let p be the initial searcher efficiency for fresh carcasses, or, more precisely, the conditional probability of detecting a carcass on the first search after carcass arrival, given that the carcass is present at the time of the search. Let k be the fractional change in searcher efficiency with each successive search. Then, if searcher efficiency trial carcasses that are missed in one search are left in the field for possible discovery on later searches, p and k can be estimated simultaneously as functions of categorical covariates.

Specifically, let p_i be the probability that carcass i is found during the first search after carcass arrival given that it is present at the time of the search. The model allows p_i to depend on a vector of covariates $\mathbf{X}_i$ as $\text{logit}(p_i) = \mathbf{X}_i\beta$, where β is a vector of coefficients associated with combinations of covariate levels. The model allows a constant multiplicative reduction in detection probability in subsequent searches. The probability of finding carcass i during search j is

$$\Pr\{\text{detect carcass } i \text{ on occasion } j\} = p_i k_i^{j-1} \tag{1}$$

where k_i may also be a logit-linear function of covariates and coefficients. These covariates need not be the same as those used to model p_i.

Consider the search outcome history for a given carcass. If the carcass was scavenged before the first search occasion, then the carcass provides no information to estimate searcher efficiency, and the model ignores these carcasses. If the carcass is present during one or more search occasions, then the probability of its search history is determined by that history and equation (1). It is easiest to see with a couple of examples. Suppose carcass i (from the searcher efficiency field trials) was missed on three searches and detected on the fourth search. Its search history is then (0, 0, 0, 1) (in the notation of the data), and the probability of this search history is

$$\Pr\{(0,0,0,1)\} = (1 - p_i)(1 - p_i k_i)\left(1 - p_i k_i^2\right) p_i k_i^3 = L_i$$

where L_i denotes the contribution of carcass i to the joint likelihood.

A carcass may never be found. For example, suppose carcass i' was missed on the first two search occasions and was then scavenged before the third search. This carcass has search history (0, 0), and

$$L_{i'} = \left(1 - p_{i'}\right)\left(1 - p_{i'}k_{i'}\right)$$

The log-likelihood of the data is then expressed as

$$\log L = \sum_{i=1}^{N} \log L_i$$

where N denotes the total number of carcasses in the trial. Let z_i denote the number of zeros in the search history for carcass i (that is, z_i is the number of times carcass i was missed in searches), and let f_i denote the search occasion where the carcass was found, with $f_i = 0$ if the carcass was never found. The, $\log L_i$ will then have the form

$$\log L_i = \sum\nolimits_{j=0}^{z_i - 1} \log\left(1 - p_i k_i^j\right) + \mathbf{1}_{f_i > 0} \log(p_i k_i^{f_i - 1})$$

where $\mathbf{1}_{f_i > 0} = 1$ if $f_i > 0$ and zero otherwise. Thus the log-likelihood depends on the data only through the number of missed searches and whether or not the carcass was eventually found.

The log-likelihood, L, of the data is numerically maximized in using R's `optim` function. From the theory of maximum likelihood, the vector of maximum likelihood estimators (MLEs) for the parameters ($\widehat{\beta}$) is asymptotically multivariate normal with mean equal to the true parameter values and variance-covariance matrix given by the inverse information matrix, which is returned by `optim` as the `hessian`.

To account for the uncertainty in estimating p and k parameters, GenEst first approximates the sampling distribution of $\widehat{\beta}$ by simulating from the multivariate normal (MVN)

$$\widehat{\beta}_{\text{sim}} \sim \text{MVN}(\widehat{\beta}, \mathbf{H}^{-1})$$

where **H** is the Hessian matrix returned by `optim()`. Simulated sampling distributions of p_i and k_i are obtained by back transformation of simulated $\widehat{\beta}$. Specifically, let $\mathbf{X}_i$ represent the covariate vector for the ith carcass, then

$$\widehat{k}_i = \frac{1}{1 + \exp(-\mathbf{X}_i \cdot \widehat{\beta}_k)}$$

$$\widehat{p}_i = \frac{1}{1 + \exp(-\mathbf{X}_i \cdot \widehat{\beta}_p)}$$

where $\hat{\beta}_p$ and $\hat{\beta}_k$ are the components of $\hat{\beta}$ associated with p and k, respectively.

8 Carcass Persistence

Carcass persistence times are modeled using censored exponential, Weibull, lognormal, and loglogistic survival models, which are fit by maximum likelihood estimation using the R functions `survival::survreg` (Therneau 2015) and `optim` (R Core Team, 2018). Both the location and scale parameters (Kalbfleisch and Prentice 2002, chapter 2) may depend on categorical covariates, such as visibility class, season, or other factor. As with the searcher efficiency estimates, vectors of simulated persistence parameters are generated from the fitted models.

References Cited

Dalthorp, D. 2018. cbinom: Continuous Analog of a Binomial Distribution. R package, version 1.3.

Dalthorp, D., Simonis, J., Madsen, L., Huso, M., Rabie, P., Mintz, J., Wolpert, R., Studyvin, J., and Korner-Nievergelt, F. 2018. GenEst: Generalised Fatality Estimator. R package, version 1.0.0.

Huso, M. M. P. and Dalthorp, D. 2014. Accounting for unsearched areas in estimating wind turbine-caused fatality. The Journal of Wildlife Management 78(2): 347-358.

Ilienko, Andreii. 2013. Continuous counterparts of Poisson and binomial distributions and their properties. Annales Univ. Sci. Budapest, Sect. Comp. 39: 137-147.

Kalbfleisch, J. D. and Prentice, R. L. 2002. The Statistical Analysis of Failure Time Data. Wiley, Hoboken, NJ.

Madsen, L., Dalthorp, D., and Huso, M. 2018. Estimating a binomial index with estimated success probability using a parametric bootstrap. Environmetrics, *in review*.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Simonis, J., Dalthorp, D., Huso, M., Mintz, J., Madsen, L., Rabie, P., Studyvin, J. 2018. GenEst User Guide. https://code.usgs.gov/ecosystems/GenEst.

Therneau T. 2015. _A Package for Survival Analysis in S_. version 2.38, <URL: https://CRAN.R-project.org/package=survival>